# Sequence variation in the human angiotensin converting enzyme

Mark J. Rieder[1], Scott L. Taylor[1], Andrew G. Clark[2] & Deborah A. Nickerson[1]

**Angiotensin converting enzyme (encoded by the gene *DCP1*, also known as *ACE*) catalyses the conversion of angiotensin I to the physiologically active peptide angiotensin II, which controls fluid-electrolyte balance and systemic blood pressure. Because of its key function in the renin-angiotensin system, many association studies have been performed with *DCP1*. Nearly all studies have associated the presence (insertion, I) or absence (deletion, D) of a 287-bp *Alu* repeat element in intron 16 with the levels of circulating enzyme or cardiovascular pathophysiologies[1–3]. Many epidemiological studies suggest that the *DCP1*D* allele confers increased susceptibility to cardiovascular disease; however, other reports have found no such association or even a beneficial effect (refs 4–7). We present here the complete genomic sequence of *DCP1* from 11 individuals, representing the longest contiguous scan (24 kb) for sequence variation in human DNA. We identified 78 varying sites in 22 chromosomes that resolved into 13 distinct haplotypes. Of the variant sites, 17 were in absolute linkage disequilibrium with the commonly typed *Alu* insertion/deletion polymorphism, producing two distinct and distantly related clades. We also identified a major subdivision in the *Alu* deletion clade that enables further analysis of the traits associated with this gene. The diversity uncovered in *DCP1* is comparable to that described for other regions in the human genome[8–11]. The highly correlated structure in *DCP1* raises important issues for the determination of functional DNA variants within genes and genetic studies in humans based on marker association.**

In terms of sequence variation, some differences were evident between the groups we studied. European-American samples had 44 varying sites, of which 4 were singletons, and African-American samples had 70 varying sites, of which 22 were singletons. The difference in the proportion of singletons between these two groups was significant ($\chi^2=7.66$, $P<0.006$). The much lower incidence of singletons in the European sample also resulted in a greater departure between estimates of $\theta=4N\mu$ obtained from the number of segregating sites and the per-site heterozygosity. This difference was quantified by Tajima's *D* statistic, which was 1.522 in the European-American sample and 0.213 in the African-American sample, and is consistent with a greater departure from equilibrium in Europe than in Africa. Despite these differences, the overall nucleotide diversity across the entire gene was not significantly different between African-Americans and European-Americans ($9.7\times10^{-4}\pm4.9\times10^{-4}$ versus $7.3\times10^{-4}\pm3.8\times10^{-4}$, respectively; Table 1).

The structured pattern of sequence variation allowed for the resolution of linkage phase and inference of haplotypes. A high degree of multi-site linkage disequilibrium was evident in a block of sites near the 5′ end of the testicular form of *DCP1* (starting near site 10,514; Fig. 1; ref. 12) in individuals who were multiply homozygous for the common variant, homozygous for the rare variant or heterozygous at all sites (Fig. 2*a*). This enabled us to infer all haplotypes, which were subsequently verified by allele-specific PCR (Fig. 2*b*,*c*; ref. 13). We identified 13 distinct haplotypes, approximately one-half of which were represented more than once, with 1 haplotype found 5 times (H1, 23% of the sampled chromosomes), another haplotype 4 times (H6, 18%) and 1 haplotype 3 times (H7, 14%). The remaining 10 haplotypes were present only once in the sampled chromosome (45% of total haplotypes). We constructed a gene tree to infer the mutational steps among haplotypes. The *Alu* insertion occurred on the long branch that separates 2 major clades along with 17 other substitutions (Fig. 3*a*). This *Alu* element was inferred to be an insertion (rather than a deletion) because its sequence has closest similarity to other human-specific *Alu* elements[14], and PCR of chimpanzee DNA revealed a fragment consistent with lack of an *Alu* (data not shown). These data suggest that the *Alu* insertion occurred after the human-chimpanzee split, but is nevertheless very old due to the significant sequence variation that has accumulated on the *Alu *I* clade. Coalescent simulations produced a maximum likelihood estimate of 1,113,000±232,000 years as the time of common ancestry of the *DCP1* haplotypes, consistent with observations in several other human nuclear genes[8,11,15,16].

A notable feature of *DCP1* sequence variation was the high degree of linkage disequilibrium in the 18 varying sites (including the *Alu* indel) that distinguish the **I* and **D* clades. We tested the hypothesis that such a pattern of SNPs might occur by chance in a

**Table 1 • Sequence diversity in different regions of *DCP1* and between ethnic groups**

| Region scanned[a] | African-American[b] (n=5) | European-American[b] (n=6) | Total[b] (n=11) |
|---|---|---|---|
| coding | 9.4±4.8 (*S*=13) | 6.1±4.0 (*S*=7) | 8.8±5.2 (*S*=15) |
| noncoding | 9.7±5.3 (*S*=57) | 7.6±4.1 (*S*=37) | 9.4±4.8 (*S*=63) |
| 5′ region | 4.1±3.3 (*S*=4) | 3.3±2.8 (*S*=2) | 3.9±3.0 (*S*=4) |
| 3′ region | 5.6±5.9 (*S*=1) | 5.1±5.5 (*S*=1) | 5.4±5.4 (*S*=1) |
| intronic | 10.6±5.8 (*S*=51) | 8.1±4.4 (*S*=33) | 10.2±5.3 (*S*=57) |
| repeat elem. | 12.3±7.5 (*S*=13) | 10.3±6.3 (*S*=9) | 11.9±6.8 (*S*=14) |
| total | 9.7±4.9 (*S*=70) | 7.3±3.8 (*S*=44) | 9.3±4.8 (*S*=78) |

[a]The total number of bases analysed for each region was: coding, 4,122 bp; noncoding, 19,948 bp; 5′ region, 2,643 bp; 3′ region, 842 bp; intronic, 16,636 bp; repeat elements, 3,305 bp; and total, 24,070 bp. [b]All values are reported as mean$\times10^{-4}\pm$s.e.$\times10^{-4}$. The total number of segregating sites in each category is given by *S*.

[1]*University of Washington, Department of Molecular Biotechnology, Box 357730, Seattle, Washington 98195, USA.* [2]*Institute of Molecular Evolutionary Genetics, Department of Biology, Pennsylvania State University, University Park, Pennsylvania 16802, USA. Correspondence should be addressed to M.J.R. (e-mail: mrieder@u.washington.edu).*

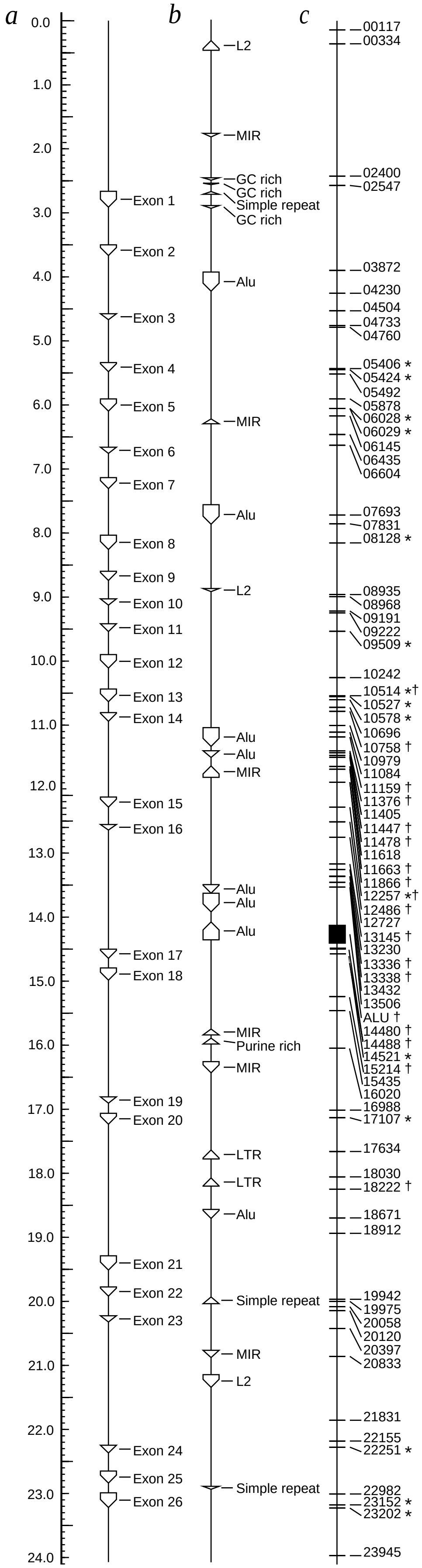

**Fig. 1** Genomic structure, repeat elements and nucleotide variations in *DCP1*. ***a***, Genomic structure of *DCP1*. *DCP1* consists of 26 exons spaced over approximately 24 kb (~16% coding sequence) and is a duplicated gene with two homologous domains (exons 1–12, 14–26). Tissue-specific expression is found in testes, with a promoter in intron 12 driving expression of the testes-specific exon 13 and the second half of *DCP1* (exons 14–26; ref. 12). ***b***, Repeat structure found in genomic *DCP1* sequence. Genomic sequence was analysed for known repetitive elements (SINE, LINE, LTR), simple repeat elements and low-complexity regions using RepeatMasker (A.F.A. Smit and P. Green; ftp.genome.washington.edu). The reverse-oriented *Alu* repeat element in intron 16 has been used as a common marker in association studies. ***c***, Variant distribution in *DCP1*. We identified 74 single-nucleotide polymorphisms (59 non-coding, 15 coding) and 4 insertion/deletion polymorphisms from 11 individuals (5 African-Americans, 6 European-Americans). All individuals were genotyped for the *Alu* insertion/deletion found in intron 16. *Sites found in coding regions (5/15 coding sites resulted in nonsynonymous changes). Nucleotide position for nonsynonymous variants are: 06029 (G/A; Ser→Ala), 09509 (C/T; Arg→Trp), 10527 (T/C; Ser→Ala), 10578 (A/G; Ser→Gly) and 23152 (C/A; Ser→Arg). Note: 2 of 5 coding variants (sites 10527 and 10578) lie in the testis-specific *DCP1* exon 13. †Variant sites in absolute linkage disequilibrium with the *Alu* insertion/deletion.

neutral gene by simulating 100,000 non-recombining neutral coalescent trees, of which 5,083 had two clades of 9 and 13. The number of trees that had 18 or more sites in absolute disequilibrium was 154, indicating that this degree of linkage disequilibrium is unlikely ($P$=0.030). This unique feature suggests several explanations. For example, it is possible that either the *Alu* polymorphism may cause destabilization by DNA mispairing, causing a flurry of additional mutations around it, or the *Alu* insertion occurred in a particular background that is rapidly increasing in frequency. Such a sweep is unlikely in an equilibrium population. Demographic factors, such as survival of a few lineages through a bottleneck or an ancestral population with a high degree of subdivision, have a much greater likelihood of causing grossly different alleles with multiple-site disequilibrium. The multiple-site linkage disequilibrium is also consistent with the apparently low levels of intragenic recombination, as assessed by the four-gamete test. Natural selection may have favoured long-term maintenance of the **I* and **D* haplotypes in the population, allowing more differences to accumulate than expected under neutrality.

Many investigators have focused only on coding regions in their search for sequence variation, arguing that these sites will exhibit the primary functional effects[10]. Had our study examined only coding regions, we would have found only 2 synonymous coding variants in the group of 17 sites in absolute disequilibrium with *Alu *I*/**D* and missed the 15 noncoding sites in this region. Numerous reports have shown an effect of noncoding variations for mendelian disorders such as cystic fibrosis[17] and phenylketonuria[18] and in model organisms such as *Drosophila melanogaster*[19]. Considering only the coding sites would also lead to a simpler interpretation of the underlying genetic structure in the population (Fig. 3*b*), although the nucleotide diversity would have been close to that obtained for the entire gene (Table 1).

Several studies have claimed to find physiological effects associated with the *Alu* insertion/deletion, including altered enzyme levels/activities, cardiovascular pathophysiologies such as myocardial infarction, arterial hypertension and left ventricular hypertrophy, elite athletic performance and response to physical training[1–4,20]. Our work shows for the first time that there is a major genetic subdivision in the deletion clade (H1 and H7) in one of the populations (European-Americans) that enables a more detailed analysis of common cardiovascular traits associated with this variation. Because the deletion form has been consistently associated with cardiovascular disease pathology, this suggests genetic heterogeneity is of potential medical importance. The generation of haplotypes from these data will also allow the application of other cladistic-based analyses to investigate genotype-phenotype relationships[21,22].

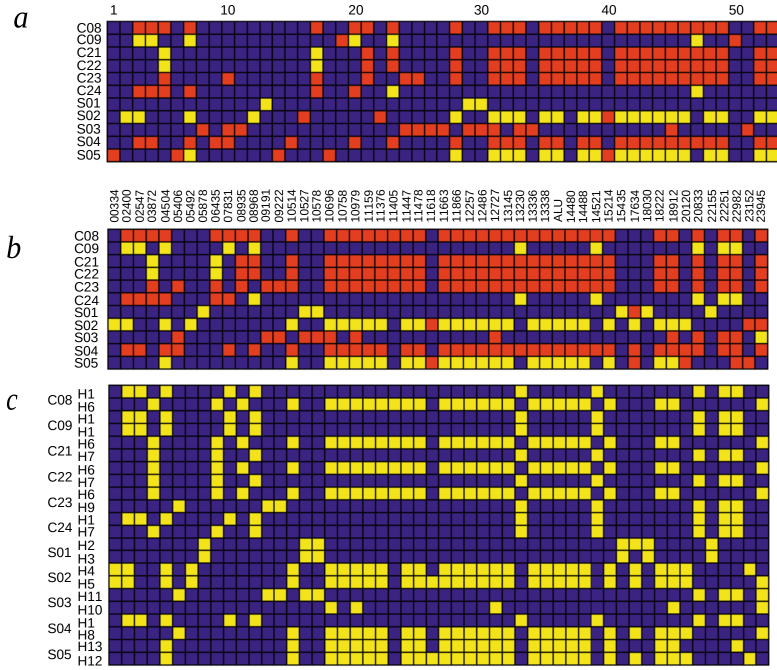

**Fig. 2** *DCP1* genotypes and haplotypes. ***a***, Genotypes of all individuals at each of the 78 variant sites. Individual sample identifiers are shown on the left side of the array (C, European-American; S, African-American) and variants are numbered consecutively across the top. Genotypes for each individual were assigned directly from the fluorescence sequencing trace at each position[24]. At each site, all individuals homozygous for the common allele are blue, heterozygotes are red and homozygotes for the rare allele are yellow. ***b***, Genotypes of all individuals at the 52 non-unique polymorphic sites (singletons excluded). ***c***, Corresponding haplotypes for each individual[13]. We resolved 13 unique haplotypes from the 22 chromosomes present. Beginning at approximately position 10514 is the testicular-specific exon 13, which also corresponds with the 5´ end of the region with a high degree of multi-site linkage disequilibrium. Blue, common allele; yellow, rare allele.

The block of sites in linkage disequilibrium highlights how difficult it may be to assign function to a particular nucleotide site. Our results suggest that associations would also be found with any of the other 17 varying sites in absolute linkage disequilibrium (or other sites simply in linkage disequilibrium) with the *Alu* element, although the extent of this disequilibrium will need to be studied in a larger sample size to fully quantify its pattern and extent in this region. Determining which (if any) of these sites may be responsible for a physiological effect will be a challenging task, especially in the absence of recombinants. The high level of disequilibrium observed in *DCP1* emphasizes the importance of obtaining a more complete picture of the allelic variation within a gene. Although the variation presented here for *DCP1* is simpler in terms of the number of variants and haplotypes than several other genes that have been analysed[8,16,23], our work, based on the number of variants present and also their linkage disequilibrium, suggests cautious interpretation of a phenotypic association with a single SNP is warranted.

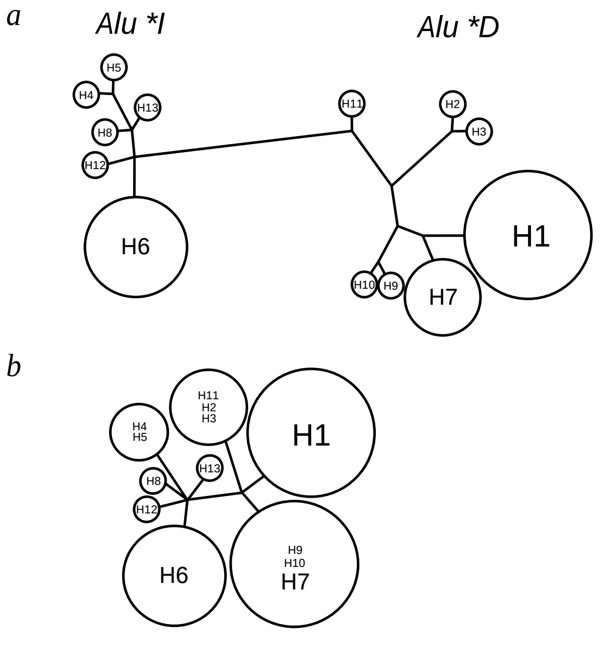

**Fig. 3** Consensus parsimony gene tree for *DCP1* haplotypes. ***a***, Gene tree constructed from the 52 non-unique variant sites. The only haplotypes that were present more than once in the sample were H1 (n=5), H6 (n=4) and H7 (n=3). The *Alu *I/*D* occurs along the long branch separating the two major clades. The relative frequency of each haplotype is indicated by the area of the circle. ***b***, A representative gene tree constructed from the eight non-unique coding sites.

## Methods

**Polymorphism detection.** European-American DNA samples (n=6) were selected from the parental generation of the CEPH (Centre d'Etude du Polymorphisme Humain) reference families (Utah and French). African-American DNA samples (n=5) were randomly selected from individual samples at the Coriell Cell Repositories. Overlapping primer sets (39) spanning the entire genomic sequence of *DCP1* were chosen on the basis of size and overlap of PCR amplicons (average size, 842±180 bp; average overlap, 260±57 bp). Before synthesis of the final primer set, either a universal forward (–21 M13, 5´–TGTAAAACGACGGCCAGT–3´) or reverse (M13 reverse, 5´–CAGGAAACAGCTATGACC–3´) sequence was added to the 5´ end of each primer to produce PCR fragments compatible with dye-primer fluorescence-based sequencing. All samples were amplified from genomic DNA (20 ng) in reactions (25 µl) using a standard PCR buffer (10 mM Tris-HCl, pH 8.3, 50 mM KCl, 1.5 mM $MgCl_2$, 0.001% gelatin, 40 mM dNTPs, 0.5 mM primer, 0.5 U of *Taq* polymerase (Perkin-Elmer Cetus)), Advantage-GC Genomic PCR kit (Clontech), or Elongase Enzyme mix (Gibco BRL). All primer sequences and PCR conditions are available (http://droog.mbt.washington.edu). PCR amplicons were sequenced and compared between all individuals to identify polymorphic sites and to genotype each individual using the PolyPhred program[24]. DNA sequence quality was used to trim all sequencing reads so only regions with an average quality of 30 or more were scanned for variations[25]. In addition, most variant sites were in regions with redundant opposite strand coverage; this amounted to more than 60% of the sequence. All sequence variants were visually inspected, scored and automatically entered into a database for subsequent analysis. Furthermore, all variants identified were confirmed by additional PCR and sequencing or allele-specific PCR analysis. All variants identified have been deposited in dbSNP and GenBank and are available (http://droog.mbt.washington.edu).

Haplotype determination. Following removal of singleton sites, haplotypes were inferred from the samples using a heuristic algorithm based on population genetic principles[13]. Direct molecular haplotyping was used to confirm a subset of heterozygous sites (~25% of 178 sites) using allele-specific PCR, allele-specific PCR combined with direct sequencing of the specific products, RFLP analysis or direct sequencing of parental samples of the CEPH individuals. All inferred haplotypes were confirmed by these molecular haplotyping techniques.

Statistical analysis. Under the infinite-sites model of molecular population genetics, the expected number of segregating sites, *S*, in an equilibrium population will be

$$\theta\sum_{i=1}^{n-1}\frac{1}{i}\ ,$$

where *n* is the sample size and $\theta = 4N\mu$, *N* being the population size and $\mu$ the neutral mutation rate[26]. Using the observed number of segregating sites to estimate $\theta$, we get $\theta_S$=21.399. Tajima's *D* statistic[27] is defined as

$$\frac{d}{\sqrt{Var(d)}}\ \text{, where }\ d=E(k)-\frac{S}{\sum_{i=1}^{n-1}\frac{1}{i}}\ ,$$

*k* is the average per-site heterozygosity and *n* is the sample size. The expected value of *D* is 0 for a neutral gene in an equilibrium population.

The program DNAPARS in Phylip 3.5 was used to infer the maximum parsimony tree for the 52 phylogenetically informative sites in the *DCP1* sequences. Five sites (2400, 3872, 4504, 6435 and 15214) were found to exhibit homoplasy and were eliminated from further genealogical analysis. For the gene tree constructed from the non-unique coding sites, DNAPARS produced 32 equally parsimonious trees. We generated a consensus tree using the gene tree constructed from all non-unique sites. Maximum likelihood estimates of $\theta=4N\mu$ and time to the most recent common ancestor ($T_{MRCA}$) were obtained using Monte Carlo Markov Chain simulations of the coalescent process, as implemented by the program Genetree[28,29]. The likelihood was significantly improved by allowing for population growth with a maximum likelihood estimate of the growth parameter $\beta$=0.16. With this growth parameter, the maximum likelihood estimate for $\theta_{ML}$ is 26.95. On the basis of an alignment of human, mouse and rat *DCP1*-coding regions, we estimate the neutral substitution rate to be $2.74\times10^{-9}$ per site per year. We used the maximum likelihood estimates of $\theta$ and $\beta$ and generated $10^7$ trees to obtain a stable estimate of $T_{MRCA}$ of 0.545±0.113 $N_e$ generations. $N_e$ was estimated from $\theta_S$=21.40 and $\mu=2.74\times10^{-9}$ to be 81,054, and from $\theta_{ML}$=26.95 to be 102,115. If one human generation is 20 years, these estimates correspond to 883,698±184,000 and 1,113,000±232,000 years, respectively. The four-gamete test was applied to infer past intragenic recombination[30]. Of 1,176 site pairs, only 114 showed the presence of all 4 gametes, and all cases involved the 5 homoplasy sites excluded from the gene tree.

Coalescence simulation. The probability of finding a block of 18 sites in absolute linkage disequilibrium by chance in a population in mutation-drift equilibrium was tested by constructing 100,000 neutral coalescent trees, retaining only those having a clade of 13 and a clade of 9. Mutations were placed on the branches of the trees following a Poisson distribution whose mean produced the observed 52 non-singleton sites, and the count of sites that were in absolute disequilibrium in the two clades was tallied.

GenBank Accession number. *DCP1*, AF118569, AC002345

**Acknowledgements.**

*We thank P. Green, K. Weiss and U. Petralia for helpful comments. This work was supported in part by HG01436 and HL58238 (D.A.N.), HL58240 (A.G.C.) and NRSA Fellowship 1 F32 HG00187-01 (M.J.R.).*

Received 11 February; accepted 30 March 1999.

1. Rigat, B. *et al.* An insertion/deletion polymorphism in the angiotensin I-converting enzyme gene accounting for half the variance of serum enzyme levels. *J. Clin. Invest.* **86**, 1343–1346 (1990).
2. Schunkert, H. Polymorphism of the angiotensin-converting enzyme gene and cardiovascular disease. *J. Mol. Med.* **75**, 867–875 (1997).
3. Soubrier, F., Nadaud, S. & Williams, T.A. Angiotensin I converting enzyme gene: regulation, polymorphism and implications in cardiovascular diseases. *Eur. Heart J.* **15**, 24–29 (1994).
4. Evans, A.E. *et al.* Polymorphisms of the angiotensin-converting-enzyme gene in subjects who die from coronary heart disease. *Q. J. Med.* **87**, 211–214 (1994).
5. O'Malley, J.P., Maslen, C.L. & Illingworth, D.R. Angiotensin-converting enzyme DD genotype and cardiovascular disease in heterozygous familial hypercholesterolemia. *Circulation* **97**, 1780–1783 (1998).
6. O'Donnell, C.J. *et al.* Evidence for association and genetic linkage of the angiotensin-converting enzyme locus with hypertension and blood pressure in men but not women in the Framingham Heart Study. *Circulation* **97**, 1766–1772 (1998).
7. Schächter, F. *et al.* Genetic associations with human longevity at the APOE and ACE loci. *Nature Genet.* **6**, 29–32 (1994).
8. Harding, R. *et al.* Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am. J. Hum. Genet.* **60**, 772–789 (1997).
9. Wang, D.G. *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082 (1998).
10. Collins, F.S., Guyer, M.S. & Charkravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science* **278**, 1580–1581 (1997).
11. Nachman, M.W., Bauer, V.L., Crowell, S.L. & Aquadro, C.F. DNA variability and recombination rates at X-linked loci in humans. *Genetics* **150**,1133–1141 (1998).
12. Hubert, C., Houot, A., Corvol, P. & Soubrier, P. Structure of the angiotensin I-converting enzyme gene. Two alternate promoters correspond to evolutionary steps of a duplicated gene. *J. Biol. Chem.* **266**, 15377–15383 (1991).
13. Clark, A.G. Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.* **7**, 111–122 (1990).
14. Batzer, M.A. *et al.* African origin of human-specific polymorphic Alu insertions. *Proc. Natl Acad. Sci. USA* **91**, 12288–12292 (1994).
15. Hey, J. Mitochondrial and nuclear genes present conflicting portraits of human origins. *Mol. Biol. Evol.* **14**, 166–172 (1997).
16. Clark, A.G. *et al.* Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.* **63**, 595–612 (1998).
17. Zielenski, J. & Tsui, L.C. Cystic fibrosis: genotypic and phenotypic variations. *Annu. Rev. Genet.* **29**, 777–807 (1995).
18. Scrivner, C.R., Byck, S., Prevost, L. & Hoang, L. in *Variation in the Human Genome* (ed. Weiss, K.M.) 73–90 (John Wiley & Sons, Chichester, 1996).
19. Long, A.D., Lyman, R.F., Langley, C.H. & Mackay, T.F. Two sites in the Delta gene region contribute to naturally occurring variation in bristle number in Drosophila melanogaster. *Genetics* **149**, 999–1017 (1998).
20. Montgomery, H.E. *et al.* Human gene for physical performance. *Nature* **393**, 221–222 (1998).
21. Sing, C.F., Haviland, M.B., Zerba, K.E. & Templeton, A.R. Application of cladistics to the analysis of genotype-phenotype relationships. *Eur. J. Epidemiol.* **8** (suppl. 1), 3–9 (1992).
22. Keavney, B. *et al.* Measured haplotype analysis of the angiotensin-I converting enzyme gene. *Hum. Mol. Genet.* **7**, 1745–1751 (1998).
23. Nickerson, D.A. *et al.* DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nature Genet.* **19**, 233–240 (1998).
24. Nickerson, D.A., Tobe, V.O. & Taylor, S.L. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res.* **25**, 2745–2751 (1997).
25. Ewing, B., Hillier, L., Wendl, M.C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**,175–185 (1998).
26. Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).
27. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphisms. *Genetics* **123**, 585–595 (1989).
28. Griffiths, R.C. & Tavaré, S. Ancestral inference in population genetics. *Stat. Sci.* **9**, 307–319 (1994).
29. Tavaré, S., Griffiths, R.C. & Donnelly, P. Inferring coalescence times from DNA sequence data. *Genetics* **145**, 505–518 (1997).
30. Hudson, R.R. & Kaplan, N.L. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164 (1985).